Automatic Detection of Machine Generated Text: A Critical Survey

Year: 2020 | Citations: 271 | Authors: Ganesh Jawahar, Muhammad Abdul-Mageed, L. Lakshmanan

Abstract

Text generative models (TGMs) excel in producing text that matches the style of human language reasonably well. Such TGMs can be misused by adversaries, e.g., by automatically generating fake news and fake product reviews that can look authentic and fool humans. Detectors that can distinguish text generated by TGM from human written text play a vital role in mitigating such misuse of TGMs. Recently, there has been a flurry of works from both natural language processing (NLP) and machine learning (ML) communities to build accurate detectors for English. Despite the importance of this problem, there is currently no work that surveys this fast-growing literature and introduces newcomers to important research challenges. In this work, we fill this void by providing a critical survey and review of this literature to facilitate a comprehensive understanding of this problem. We conduct an in-depth error analysis of the state-of-the-art detector and discuss research directions to guide future work in this exciting area.